PREDICTING HOUSE PRICES IN KING COUNTY

BY CAROLYNE W. KAMBURA
PHASE_2 PROJECT
MORINGA SCHOOL

BUSINESS UNDERSTANDING

- From observations and experience, one of the greatest challenge when searching for a house to buy is finding a suitable property within your budget range.

- It can be really frustrating when price becomes the determinant factor of whether you get to own your dream house or not.

- The current market trends show that house prices are very volatile and have not been well standardized, therefore, sellers get to set apply price based on their needs and market rates based on location regardless of size and quality.

- So this prompts the big question:

****** What factors influence property value?******

PROJECT OVERVIEW

Hypothesis:

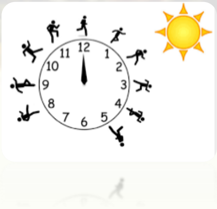
There are more than a single factor that influence house pricing.

Objectives:

- ✓ Provide Insight to homebuyers on key factors that influence price
- ✓ Develop statistical models that help predict house prices
- ✓ Understand the relationship between price and different factors assumed to influence market prices

Project Scope:

In this case study, we explored housing sale prices in King County(KC), Seattle USA to help answer our hypothesis. The dataset includes house prices and different factors that are assumed to influence the prices. We analysed each feature to establish relationship with price. We then used the results to develop price prediction models.

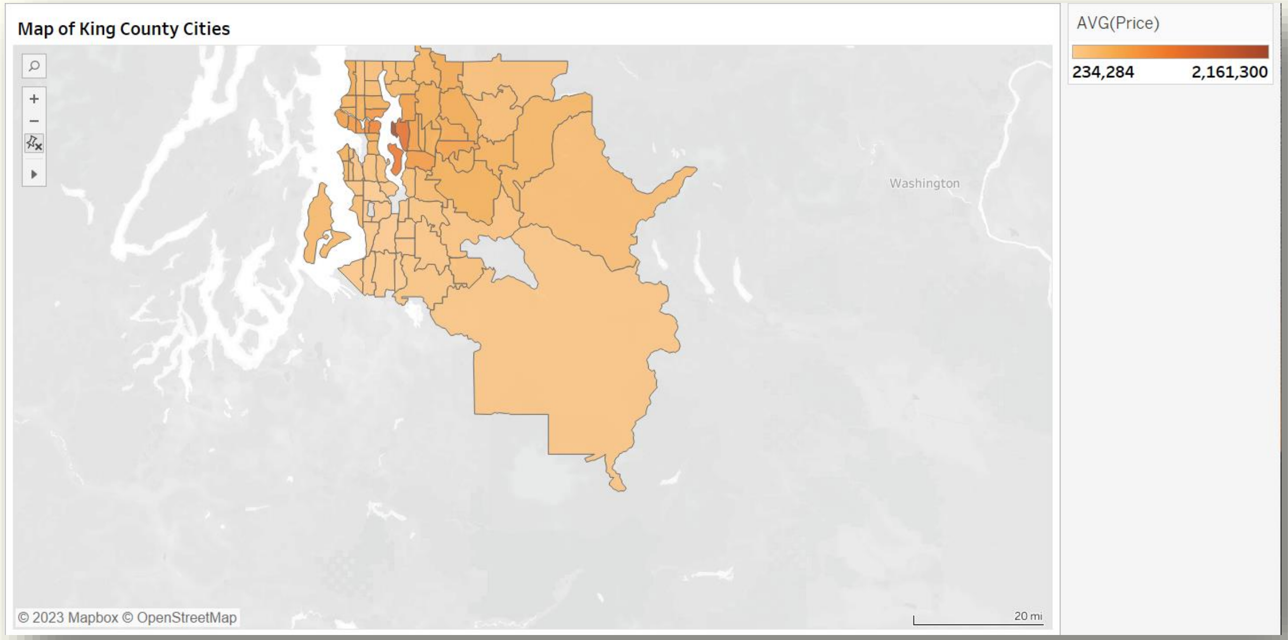


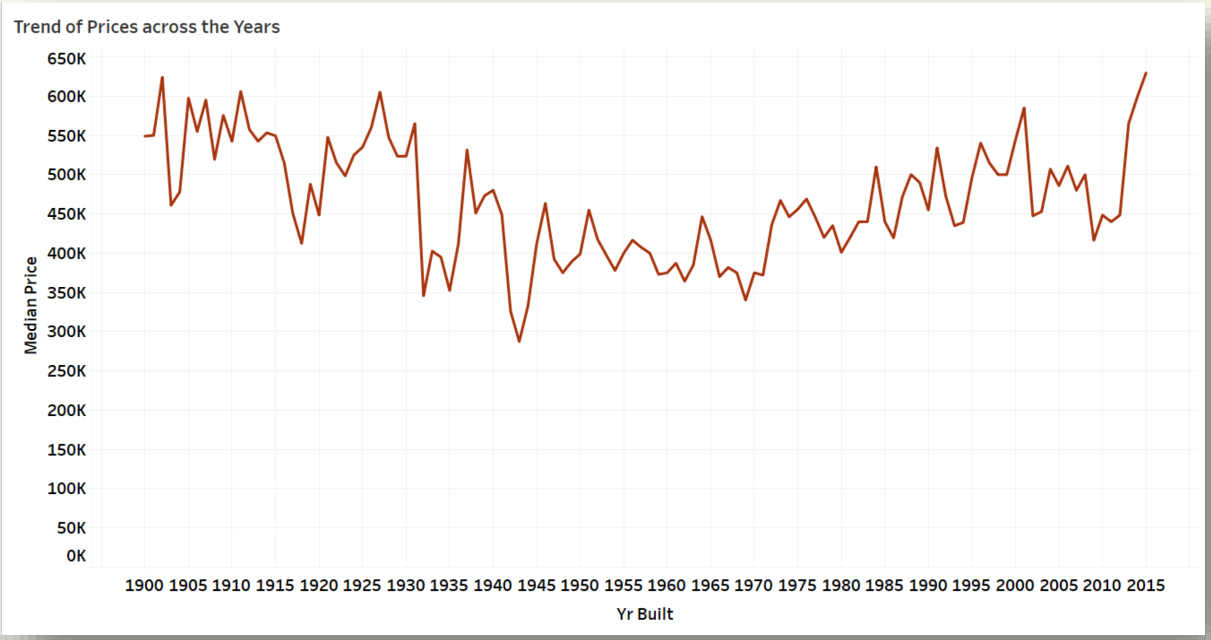
Project Start Time: 20th March 2023
Project End Time: 26th March 2023
Duration: 7 days

DATA UNDERSTANDING

General Overview KC Average House Prices

Trend of prices in KC over the years





On the trend of prices map, we used median prices from the 70 given zip codes. It is observed that there is no linear relationship between year built and prices. It is expected as house prices are volatile and are influenced by different direct and indirect factors including economic and political. (Each year cannot be the same)!

Data Source: https://www.kaggle.com/datasets/harlfoxem/housesalesprediction

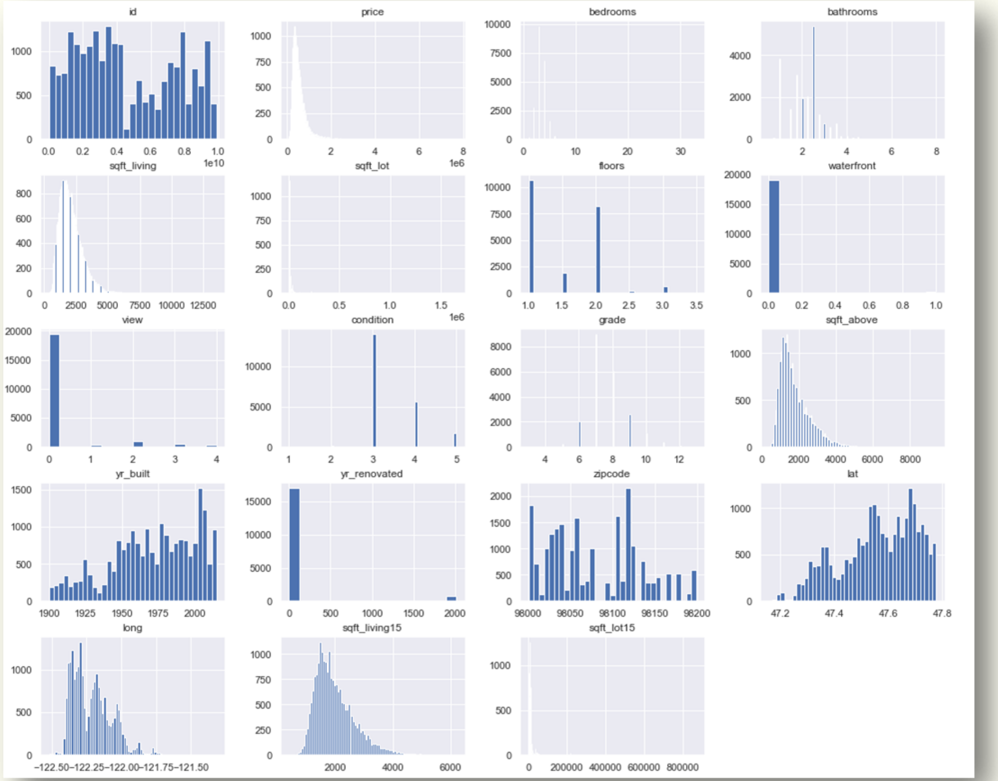
Data Overview:

Data contains :

- 20 variables (Objects, floats, & integers).
- Continuous and categorical data
- Missing values

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21597 entries, 0 to 21596
Data columns (total 21 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   id             21597 non-null  int64
 1   date           21597 non-null  object
 2   price          21597 non-null  float64
 3   bedrooms       21597 non-null  int64
 4   bathrooms      21597 non-null  float64
 5   sqft_living    21597 non-null  int64
 6   sqft_lot       21597 non-null  int64
 7   floors         21597 non-null  float64
 8   waterfront     19221 non-null  float64
 9   view           21534 non-null  float64
 10  condition      21597 non-null  int64
 11  grade          21597 non-null  int64
 12  sqft_above     21597 non-null  int64
 13  sqft_basement  21597 non-null  object
 14  yr_built       21597 non-null  int64
 15  yr_renovated   17755 non-null  float64
 16  zipcode        21597 non-null  int64
 17  lat            21597 non-null  float64
 18  long           21597 non-null  float64
 19  sqft_living15  21597 non-null  int64
 20  sqft_lot15     21597 non-null  int64
dtypes: float64(8), int64(11), object(2)
memory usage: 3.5+ MB
```

Distribution



#Observations
price, sqft_lot, sqft_living, sqft_above, long,sqft_living15 and sqft_lot15 are continuous variables and appear to be #log normally distributed. # Most houses have 3 bedrooms and 2 bathrooms
#Most houses were built in the early 2000s

NB: To be able to build models using the given data, we need to do handle outliers, perform data Standardization, normalize and deal with the categorical data

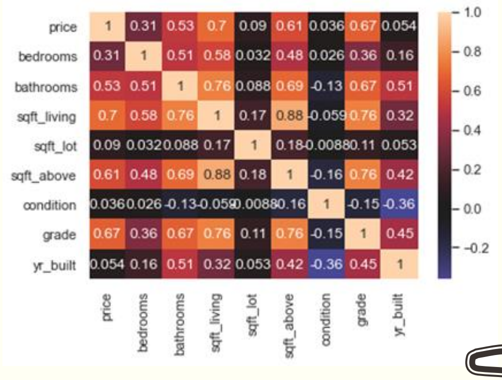
DATA PREPARATION

LINEARITY TEST

Use pairplots hue = bedrooms

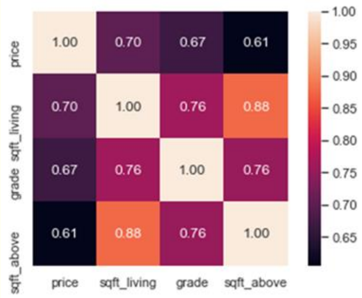
Correlation Analysis

Set correlation mark as > 0.70



Correlation to price

Identify top 3 features correlated to price

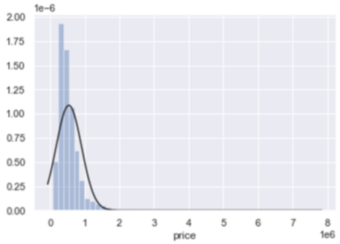


NB: Multicollinearity shows Bathrooms and sqft_living are highly correlated so we will drop “bathrooms” during modelling.

	cc
pairs	
(bathrooms, sqft_living)	0.76

Data handling

- Data normalization
-Data standardization
- Dealing with categorical data




```
#Analyse "price" for skweness and kurtosis
print("Skewness: %f" % df['price'].skew())
print("Kurtosis: %f" % df['price'].kurt())
```

```
Skewness: 4.023365
Kurtosis: 34.541359
```


```
df['price'] = np.log(df['price'])
```

DATA MODELLING

1. Simple Linear Regression

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.455
Model:                            OLS   Adj. R-squared:                  0.455
Method:                 Least Squares   F-statistic:                 1.805e+04
Date:                Sun, 26 Mar 2023   Prob (F-statistic):               0.00
Time:                        12:44:22   Log-Likelihood:                -10231.
No. Observations:               21597   AIC:                         2.047e+04
Df Residuals:                   21595   BIC:                         2.048e+04
Df Model:                           1
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
Intercept       6.7234      0.047    142.612      0.000       6.631       6.816
sqft_living     0.8376      0.006    134.368      0.000       0.825       0.850
==============================================================================
Omnibus:                      123.577   Durbin-Watson:                   1.977
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              114.096
Skew:                           0.143   Prob(JB):                     1.68e-25
Kurtosis:                       2.787   Cond. No.                         137.
==============================================================================
```

Add Sqft_lot

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.463
Model:                            OLS   Adj. R-squared:                  0.463
Method:                 Least Squares   F-statistic:                     9303.
Date:                Sun, 26 Mar 2023   Prob (F-statistic):               0.00
Time:                        08:59:32   Log-Likelihood:                -10081.
No. Observations:               21597   AIC:                         2.017e+04
Df Residuals:                   21594   BIC:                         2.019e+04
Df Model:                           2
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
sqft_living     0.8746      0.007    133.555      0.000       0.862       0.887
sqft_lot       -0.0534      0.003    -17.330      0.000      -0.059      -0.047
const           6.9238      0.048    143.563      0.000       6.829       7.018
==============================================================================
Omnibus:                       94.629   Durbin-Watson:                   1.977
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               92.481
Skew:                           0.143   Prob(JB):                     8.28e-21
Kurtosis:                       2.856   Cond. No.                         218.
==============================================================================
```

R-Squared of 45% means the model is only able to account for 45% of the total variation in the dependent variable, while the remaining 55% is due to other factors not included in the model or random error.

From the summary, the model is only able to account for a total variation of 46% in the dependent variable while the rest 54% remain unaccounted for. However the model has improved by 1% with the addition of sqft_lot

Model Accuracy = 45%

```
accuracy = regressor.score(x_test, y_test)
"Accuracy: {}%".format(int(round(accuracy * 100)))
```

```
'Accuracy: 45%'
```

	id	price	bedrooms	sqft_living	sqft_lot	floors	condition	grade	predictions
0	7129300520	12.31	3	7.07	8.64	1	3	7	12.65
1	6414100192	13.20	3	7.85	8.89	2	3	7	13.32
2	5631500400	12.10	2	6.65	9.21	1	3	6	12.25
3	2487200875	13.31	4	7.58	8.52	1	5	7	13.10

*******Prediction column shows price prediction**

```
print(results.predict([7.07,8.64,1]))
```

```
[12.64642298]
```

Multiple Regression

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.589
Model:                            OLS   Adj. R-squared:                  0.588
Method:                 Least Squares   F-statistic:                     1065.
Date:                Sun, 26 Mar 2023   Prob (F-statistic):               0.00
Time:                        09:05:52   Log-Likelihood:                -7195.4
No. Observations:               21597   AIC:                         1.445e+04
Df Residuals:                   21567   BIC:                         1.469e+04
Df Model:                          29
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
sqft_living     0.5757      0.011     53.716      0.000       0.555       0.597
sqft_lot       -0.0594      0.003    -19.609      0.000      -0.065      -0.053
bedrooms_2     -0.0878      0.026     -3.416      0.001      -0.138      -0.037
bedrooms_3     -0.2491      0.026     -9.664      0.000      -0.300      -0.199
bedrooms_4     -0.2730      0.026    -10.308      0.000      -0.325      -0.221
bedrooms_5     -0.2455      0.028     -8.790      0.000      -0.300      -0.191
bedrooms_6     -0.2288      0.034     -6.760      0.000      -0.295      -0.162
bedrooms_7     -0.2502      0.061     -4.083      0.000      -0.370      -0.130
bedrooms_8     -0.0957      0.098     -0.979      0.328      -0.287       0.096
bedrooms_9     -0.0039      0.141     -0.028      0.978      -0.280       0.272
bedrooms_10    -0.0676      0.197     -0.343      0.732      -0.454       0.319
bedrooms_11    -0.1599      0.339     -0.472      0.637      -0.824       0.505
bedrooms_33     0.1002      0.339      0.296      0.767      -0.564       0.765
floors_2       -0.1029      0.006    -16.886      0.000      -0.115      -0.091
floors_3       -0.0080      0.015     -0.520      0.603      -0.038       0.022
condition_2    -0.0913      0.068     -1.341      0.180      -0.225       0.042
condition_3    -0.0060      0.063     -0.095      0.924      -0.130       0.118
condition_4     0.0775      0.063      1.223      0.221      -0.047       0.202
condition_5     0.2046      0.064      3.211      0.001       0.080       0.330
grade_4        -0.2120      0.345     -0.615      0.538      -0.887       0.463
grade_5        -0.2067      0.339     -0.609      0.542      -0.872       0.458
grade_6        -0.0728      0.339     -0.215      0.830      -0.737       0.592
grade_7         0.0928      0.339      0.274      0.784      -0.572       0.757
grade_8         0.2920      0.339      0.861      0.389      -0.373       0.957
grade_9         0.5477      0.339      1.615      0.106      -0.117       1.213
grade_10        0.7766      0.339      2.288      0.022       0.111       1.442
grade_11        1.0039      0.340      2.955      0.003       0.338       1.670
grade_12        1.2825      0.341      3.758      0.000       0.614       1.951
grade_13        1.6084      0.352      4.566      0.000       0.918       2.299
const           9.2338      0.350     26.384      0.000       8.548       9.920
==============================================================================
Omnibus:                       34.523   Durbin-Watson:                   1.976
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               35.194
Skew:                           0.086   Prob(JB):                     2.28e-08
Kurtosis:                       3.096   Cond. No.                     5.80e+03
==============================================================================
```

Great we see that the model does improve to R-squared to 59% when we add in the categorical data. We also observe that the data is symmetrical with a very slight tendency towards the right but it's very small to be significant.

Model Validation

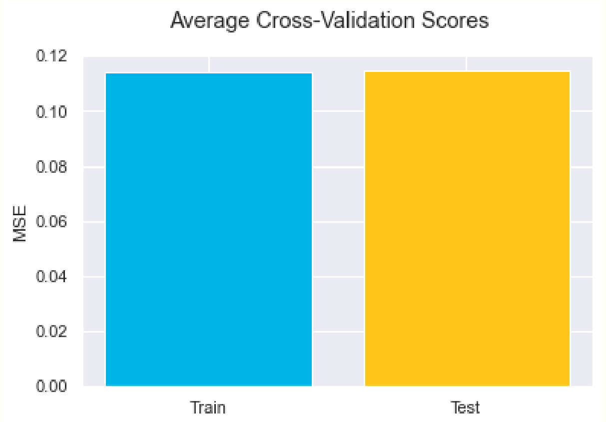
Split Train Test

Results:

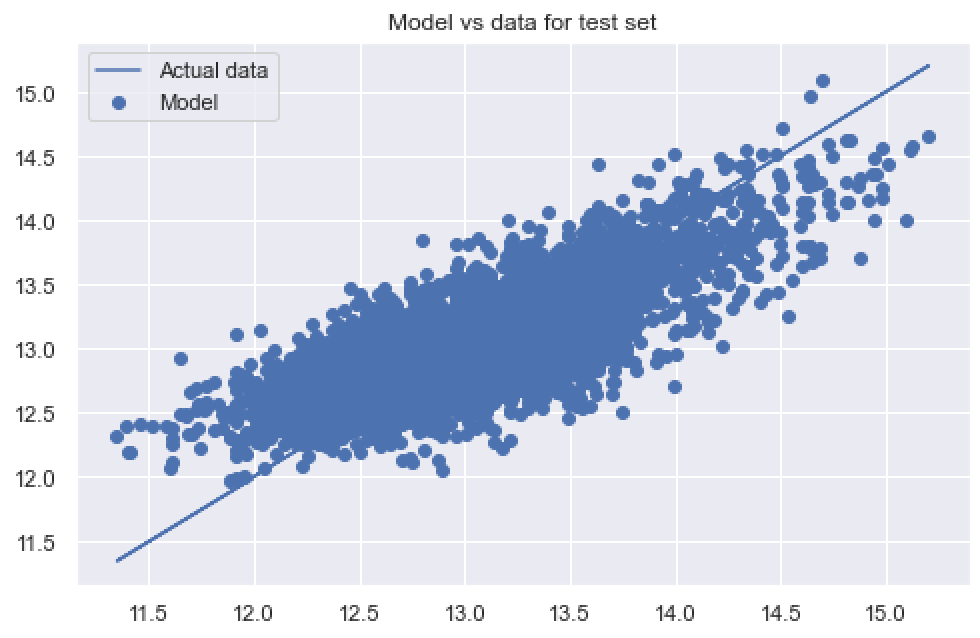
```
Train Mean Squared Error: 0.1138816185667106
Test Mean Squared Error:  0.11433987779287012
```

The test error is not that significantly different from the train meaning that the model is able to generalize future cases well

K-Fold Cross Validation



There is no significance differece between Train_score and Test_score, the model is able to generalize future cases



We observe that the data has Homoscedasticity i.e. dependent variable is equal across values of the independent variables

Bias-Variance Tradeoff:

Results:

```
Train bias: 4.134633132394236e-17
Train variance: 0.16304818835004653
```

```
Test bias: 0.0013848834907673557
Test variance: 0.1605880164739517
```

*******From the results, our model has a relatively low bias and variance, therefore predictions will be accurate

Conclusion & Recommendation

Conclusions:

- The Simple linear regression model can only account for 45% of the total variation for the dependent variables
- Multiple regression improves the model as we are now able to account for 59% of the total variations for the dependent variables
- Both model validation methods i.e. split trai-test and K-Fold shows that there is no significant difference between the actual and model data.
- Bias-Variance tradeoff shows relatively low bias and variance.
- We fail to reject the null hypothesis.

Recommendation:

- Apply other advanced methods to see if this improves predictions